LECTURA RECOMENDADA

DATOS DE INVESTIGACIÓN

Datos de investigación

- Datos de investigación, datos FAIR y gestión de datos
- Localización y citación de datos
- Herramientas de análisis y visualización de datos
- Depósito de datos

Índice rápido

Nota de los formadores:

En esta guía se abordan algunos de los elementos más importantes en la gestión de datos de investigación, incluyendo una introducción a los datos FAIR, fuentes para encontrar y publicar datos, así como herramientas para analizar y visualizar datos de la investigación.

Introducción

Los rápidos avances en computación y el desarrollo de la ciencia de datos en las últimas décadas han cambiado radicalmente la forma en que se lleva a cabo la investigación. Los **datos de investigación** son hoy en día un elemento central en el proceso científico y cada vez cobran más atención y relevancia.

Además, en el marco de la **Ciencia Abierta,** las instituciones e investigadores están impulsando la publicación en abierto de datos de investigación para su reutilización e intercambio y así eliminar barreras de acceso y acelerar la generación del conocimiento.

La Comisión Europea fue una de las primeras instituciones en promover la publicación de **datos en abierto** mediante su programa Horizonte 2020, hoy Horizonte Europa, que requiere la elaboración de un **Plan de Gestión de Datos** para todos los proyectos financiados con estos fondos, así como su publicación en abierto, respetando las restricciones de la legislación de protección de datos personales y los derechos de autoría, siguiendo el lema ***tan abiertos como sea posible, tan cerrados como sea necesario***.

En España las instituciones han comenzado también a impulsar la publicación de datos de investigación, como queda reflejado en las recientes Ley Orgánica del Sistema Universitario (LOSU, 2023), la modificación de la Ley de la Ciencia, la Tecnología y la Innovación (LCTI, 2011 modificada 2022) o la Estrategia Nacional de Ciencia Abierta (ENCA, 2023).

> Los datos, entendidos como aquellas fuentes primarias necesarias para validar los resultados de las investigaciones, deberán seguir los principios FAIR (datos fáciles de encontrar, accesibles, interoperables y reutilizables) y, siempre que sea posible, difundirse en acceso abierto. (LCTI. Artículo 37)

La iniciativa europea Coalition for Advancing Research Assessment (COARA, 2022) defiende además la necesidad de considerar los datos de la investigación como un resultado de investigación con el mismo valor que el artículo científico. Es decir, que los datos puedan ser publicados, citados y utilizados en la evaluación de la investigación.

Datos de investigación

Los **datos de investigación** son nuevas fuentes de información que permiten la reproducibilidad de los análisis, mejorar la comprensión y alcance de los resultados y contribuir a acelerar los descubrimientos a través de la reutilización de los datos.

No existe una única definición de datos de investigación, puesto que estos varían enormemente entre las distintas disciplinas con diferentes formatos, medios, etc. No obstante, en la ENCA se incluye la siguiente definición:

> *Los datos de investigación son todo aquel material que ha sido generado, recopilado, observado o registrado durante el ciclo de vida de un proyecto de investigación y que se utiliza como evidencia de un proceso de investigación, está reconocido por la comunidad científica y sirve para validar los resultados de la investigación y garantizar su reproducibilidad.*

En esta guía se va a utilizar el concepto de datos de la investigación en dos sentidos:

a) Materiales recopilados o recolectados para llevar a cabo la investigación: encuestas, fuentes de datos públicas, fotografías, etc.
b) Materiales desarrollados durante la propia investigación y materiales que apoyen la reproducibilidad de la ciencia: resultados del análisis computacional de los datos, cuadernos de laboratorio, código, etc.

En la sociedad del conocimiento los datos normalmente se presentan en un formato digital: documentos de texto, hojas de cálculo, imágenes, gráficos, bases de datos, información georreferenciada, etc., sin olvidar los datos analógicos.

Los datos deben ser interpretados para que proporcionen información y ayuden a adoptar decisiones apoyadas en datos.

Tipos de datos

Los datos pueden ser definidos según su forma, análisis, etc.

Sin ánimo de ser exhaustivos, algunas de estas clasificaciones son:

1. Según el nivel de accesibilidad: abiertos o privados.
2. Según su organización: datos estructurados o no estructurados.
3. Según su origen: experimentales, observacionales, simulación, etc.
4. Según su obtención: primarios o secundarios.
5. Según su estado: en bruto, procesados, anonimizados, etc.
6. Según su procesamiento computacional: números enteros, números flotantes, cadenas de textos, booleanos, etc.
7. Según su formato de almacenamiento: JSON, CSV, PNG, WAV, etc.

Algunos ejemplos de datos de investigación podrían ser:

- Textos (entrevistas, guías, protocolos, metodologías, etc.).
- Datos numéricos (respuestas a test, hojas de cálculo, información geoespacial, etc.).
- Diapositivas, diseños y muestras.
- Fotografías e imágenes.
- Cortes de películas o vídeos.

- Registros sonoros.
- Programas de software o código utilizado para generar o analizar los datos.
- Algoritmos.
- Desarrollo de modelos.
- Cuadernos de laboratorio.
- Cuadernos de campo.
- Muestras biológicas.
- Colecciones de objetos físicos.

Además, los datos pueden clasificarse en:

- **Datos categóricos o nominales.** Son datos cualitativos usados para nombrar una categoría. No tienen un orden o jerarquía definida. Por ejemplo:
  - Nombre
  - Color de pelo
  - Género
  - Estado civil
- **Datos ordinales.** Son datos cualitativos que se usan para ordenar valores posibles en una categoría. Se pueden ordenar y clasificar. Por ejemplo:
  - Nivel de estudios
  - Nivel de satisfacción
  - Posición en una maratón
  - Grado de dolor de un paciente
- **Datos cuantitativos discretos**. Son datos numéricos que pueden tomar una serie de valores limitada y contable. Muchas veces son números enteros. Por ejemplo:
  - Número de hijos
  - Cantidad de veces que una palabra aparece en un texto
  - Número de libros prestados en la biblioteca
- **Datos cuantitativos continuos.** Son datos numéricos que pueden tomar una serie de valores ilimitada y continua. Muchas veces presentan decimales. Por ejemplo:
  - Temperatura
  - Duración de un vídeo
  - Peso

Datos tabulares

Una de las formas más comunes de presentar los datos de investigación es la llamada matriz de datos, que presenta los datos organizados de forma tabular. Este formato se usa, puesto que su estructura es fácilmente comprensible y editable.

Los datos tabulares se presentan en tablas en formato CSV (valores separados por comas), en las que los casos (1, 2, 3, etc.) se muestran en filas y las variables (estado civil, edad, nivel de estudios, etc.) en columnas (A, B, C, etc.).

Es importante ser consistentes en la organización de datos en casos y variables para permitir su procesamiento computacional. Los datos bien estructurados se conocen como datos organizados o **tidy data.**

CSV es un formato abierto para el intercambio de datos en el contexto de la ciencia abierta. No obstante, la simplicidad de este formato tiene contrapartidas, como la ausencia de metadatos descriptivos de estos datos. De modo que conviene vincular el fichero CSV con los esquemas de metadatos que documentan y

facilitan la reutilización de los datos (fichero readme.txt, diccionarios de datos, libros de códigos, etc.).

Puede consultarse la Guía práctica para la publicación de datos tabulares en archivos CSV para obtener más información sobre este tipo de datos.

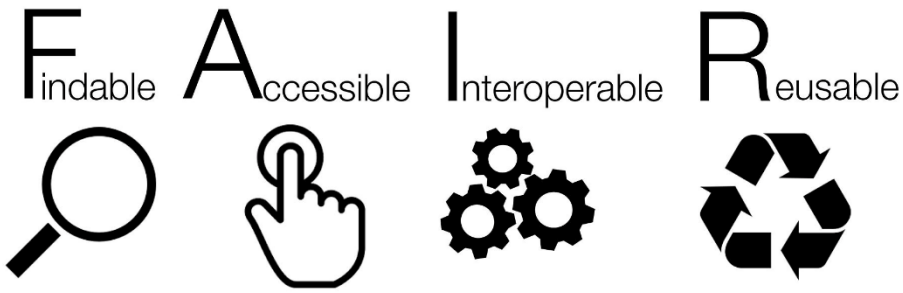
Conjuntos de datos o *datasets*

La totalidad de los datos relacionados a un proyecto de investigación se llama **conjunto de datos** (*dataset* en inglés), que incluyen además una serie de documentos que ayudan a la interpretación y utilización de los datos. Los *datasets* pueden depositarse en repositorios de datos para que puedan ser reutilizados por terceros y citados de la manera correspondiente.

Algunas de las ventajas de publicar datos en abierto son:

- Permite la reproducibilidad de la ciencia.
- Se reduce la duplicidad en la recogida de los datos.
- Se promueve la transparencia y las buenas prácticas.
- Impulsa el debate científico y la innovación.
- Mejora la visibilidad del proyecto.
- Se alinea con políticas europeas y nacionales de ciencia abierta.
- Cumplir con los requisitos de algunas convocatorias.

Datos FAIR



SangyaPundir, CC BY-SA 4.0, via Wikimedia Commons

Tanto la Comisión Europea como las instituciones españolas mencionan la necesidad, no solo de gestionar y publicar los datos, si no de que nuestros datos sean FAIR. Este acrónimo inglés implica tomar medidas específicas para que nuestros datos puedan ser encontrados y reutilizados por la comunidad científica.

Según OpenAire, algunas de las medidas que necesitan tomarse para que nuestros datos sean FAIR son las siguientes:

- **F — *Findable* o fáciles de encontrar:**
  - El conjunto de datos tiene un identificador persistente, normalmente un DOI o handle.
  - El conjunto de datos tiene los metadatos necesarios para su localización y uso.
  - El conjunto de datos puede ser encontrado en agregadores y plataformas de descubrimiento.

- **A — *Accessible* o accesibles:**
  - El conjunto de datos está depositado en un repositorio de datos.
  - El conjunto de datos está lo más abierto posible.
  - Los metadatos del conjunto de datos tienen una licencia CC-0 o CC-BY.
- **I — *Interoperable* o interoperables:**
  - El conjunto de datos sigue estándares y normas.
  - El conjunto de datos usa formatos abiertos.
- **R — *Reusable* o reutilizables:**
  - El conjunto de datos está bien documentado, con archivos *readme*, diccionarios de datos, etc.
  - El conjunto de datos tiene una licencia clara, preferiblemente una licencia Creative Commons.

Gestión de datos

En la actualidad, los proyectos de investigación en ciencia de datos se basan en el análisis de cientos —o miles— de datos de investigación. El gran volumen de datos, unidos a la necesidad de publicarlos y de manera que cumplan los principios FAIR, hace que sea necesario poner especial atención a la **gestión de datos**. Es más, entidades financiadoras como la Comisión Europea en su convocatoria Horizonte Europa, así como cada vez más convocatorias nacionales exigen al equipo investigador la elaboración de un **Plan de Gestión de Datos (PGD)**, en inglés Data Management Plan (DMP).

El Plan de Gestión de Datos (PGD) es un documento en el que se detalla la gestión de los datos desde el inicio del proyecto hasta su publicación o depósito. Para ello, es necesario planificar cada uno de los pasos que habrá que tener en cuenta durante el ciclo de vida de los datos y asegurando su vida a largo plazo una vez se concluya el proyecto.

Según la CRUE, las etapas del ciclo de vida de los datos que deben ser incluidos en el PGD son al menos:

- **Identificación de los datos:** tipología, procedencia, volumen, formatos y ficheros.
- **Organizarán y gestión de los datos:** nombre de los ficheros, control de versiones, software necesario...

- **Documentación de los datos:** información a procesar, estándares o esquemas de metadatos, herramientas, etc.
- **Calidad de los datos:** Descripción de procesos que aseguran una buena calidad de los datos y corrección de errores
- **Estrategia de almacenamiento y de preservación de datos.**
- **Políticas de datos**: cuestiones sobre propiedad intelectual, datos sensibles y personales.
- **Difusión de datos:** medios de difusión, depósito en repositorio, publicación de *data paper,* etc.
- **Roles y responsabilidades** para las personas y organizaciones participantes en el proyecto.
- **Presupuesto realista:** gastos derivados del software, hardware, servicios y personal

El PGD es un documento abierto que puede ser modificado en el transcurso del proyecto.

La redacción del PGD es libre, pero existen plantillas y herramientas específicas para la creación de un PGD. La herramienta más completa es PGDOnline.

La biblioteca de la UNED puede ayudarte en la elaboración de un PGD.

PGDOnline

Aplicación en línea desarrollada por el Consorcio Madroño para elaborar planes de gestión de datos según las directrices de Horizonte Europa, guiando paso a paso en la redacción de cada uno de los aspectos. La aplicación es una adaptación de la herramienta Argos de OpenAire (en inglés).

En el portal InvestigaM puedes consultar recursos útiles y material de apoyo.

Ciclo de vida de los datos. Fuente: RDMLA

Fuentes para localizar datos

Además de generar datos, el personal investigador puede encontrar y reutilizar numerosos *datasets* que hayan sido publicados en abierto.

Agregadores de búsqueda de datos

En la actualidad, existen infinidad de portales y repositorios de datos, por lo que puede ser conveniente usar **agregadores de búsqueda** como los siguientes:

OpenAire Explore

Iniciativa de la Comisión Europea para proporcionar las infraestructuras necesarias para el desarrollo de la Ciencia Abierta. Permite la búsqueda de *datasets* recolectados de numerosos repositorios de datos como Zenodo, BASE, o e-cienciaDatos, entre otros.

DataCite Commons

Su objetivo es ayudar a localizar, identificar y citar datos. Además, proporciona identificadores persistentes (DOI).

Figshare

Repositorio de datos de Digital Science.

DataMED

Buscador de *datasets* y repositorios en el ámbito biomédico.

DataONE

Buscador de *datasets* en repositorios sobre medio ambiente.

Mendeley Data

Buscador de datos de investigación de Elsevier.

Google Dataset Search

Servicio de Google dedicado a la búsqueda de conjuntos de datos.

También pueden localizarse conjuntos de datos en los repositorios mencionados en la sección *Depósito y publicación de datos* o encontrando repositorios temáticos a través de Re3data.

Portales de datos

Cabe mencionar también los **portales de datos de organismos e instituciones públicas** para obtener datos estadísticos. Algunos ejemplos son:

- Iniciativa Aporta: portal de datos del Gobierno de España (datos.gob.es)
- Eurostat: portal de datos estadísticos de Eurostat
- UNdata: portal de datos de las Naciones Unidas
- OECD Data: portal de datos de la OCDE
- The World Bank Open Data: portal de datos del Banco Mundial
- data.gov.uk: portal de datos del Reino Unido

Para encontrar más portales de datos en abiertos, puede usarse la plataforma de búsqueda de portales Dataportals.

European Open Science Cloud

La Comisión Europea está impulsando la **plataforma European Open Science Cloud (EOSC),** que pretende convertirse en un punto de acceso universal para servicios de datos, incluyendo algunos ya mencionados, y muchos más. Actualmente, está en fase de desarrollo.

Repositorios de código

Para el caso específico de la **programación y la búsqueda de código**, las herramientas más utilizadas son:

Kaggle

Plataforma de Google para encontrar y publicar conjuntos de datos, código, modelos, etc., especialmente orientado a la creación de modelos de aprendizaje automático e inteligencia artificial.

GitHub

Plataforma para almacenar código fuente de programas de ordenador, especialmente diseñado para el control de versiones y la colaboración.

Citar datos

Cuando se utilizan datos de terceros, es imprescindible citar la referencia de la fuente de los datos. Citar correctamente datos y *datasets* es importante por los siguientes motivos:

- Identifica al creador de los datos reconociendo su trabajo.
- Permite su reutilización, así como replicar y verificar los análisis propuestos.
- Permite seguir el impacto de los datos.
- Da visibilidad y difusión al tema de investigación.

Para la correcta citación de un conjunto de datos, los metadatos que deben identificarse son los siguientes:

- Autor(es): puede ser un autor único, múltiple o institucional
- Fecha: año de publicación de la versión.
- Título del *dataset.*
- Identificador único persistente (DOI o handle).
- Repositorio.
- Versión y/o edición.

Dependiendo del estilo de citación usado, la referencia tendrá un aspecto u otro. Los manuales de IEEE y APA tienen apartados específicos para citar *datasets,* pero otros como Chicago o MLA no. En estos casos, se recomienda citar los datos como si fueran un recurso web. DataCite ha creado la herramienta DOI Citation Formatter que permite previsualizar la referencia deseada en numerosos estilos de citación al introducir un DOI válido.

Referencias en APA

Las referencias en APA siguen el siguiente esquema:

Apellidos del autor, inicial. (año). *Título en cursiva* (versión) [Data set]. Repositorio. DOI

O'Donohue, W. (2017). *Content analysis of undergraduate psychology textbooks* (ICPSR 21600; Version V1) [Data set]. ICPSR. https://doi.org/10.3886/ICPSR36966.v1

Más información: Guía de citas en APA

Referencias en IEEE

Las referencias en IEEE siguen el siguiente esquema:

[#] Iniciales del autor. Apellido, "Título del dataset", repositorio, fecha. doi: DOI

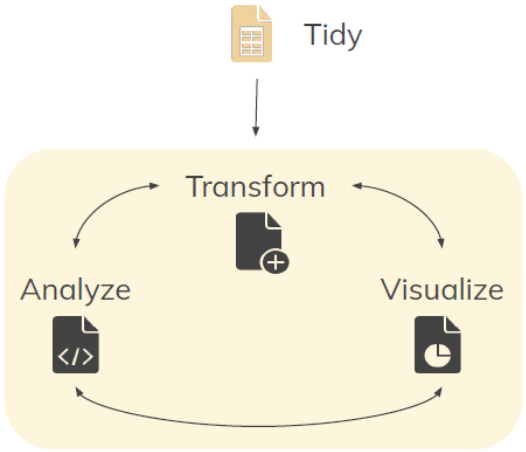
[1] W. O'Donohue, "Content analysis of undergraduate psychology textbooks (ICPSR 21600; Version V1)", ICPSR, 2017. doi: https://doi.org/10.3886/ICPSR36966.v1

Más información: IEEE Reference Guide

Herramientas para el análisis y visualización de datos

El tratamiento de datos es una cuestión compleja, puesto que depende enormemente del tipo de datos y de los objetivos del proyecto en concreto, e incluye numerosas tareas como **refinar, transformar,** analizar y visualizar los datos.

Existen numerosas herramientas que nos ayudan con cada uno de estos procesos. En esta guía se mencionan algunas de las más comunes, pero se pueden descubrir más herramientas consultando la guía de herramientas de procesado y visualización de datos de la Iniciativa Aporta.



Diferentes tareas en el trabajo con datos. Fuente: RDMLA

Herramientas de limpieza de datos:

OpenRefine

Open Refine

Aplicación de código abierto diseñada para el refinamiento de datos, con una apariencia similar a una hoja de cálculo. Entre sus funcionalidades principales, destaca:

- Limpieza de datos: unificar campos, editar grupos de celdas, etc.
- Transformación de datos: dividir columnas, combinar datos para crear nuevas variables, aplicar expresiones regulares, etc.
- Enriquecer datos a través de fuentes externas

Trabaja con numerosos formatos, incluidos JSON, XML, XLS, o CSV, entre otros.

La herramienta cuenta con un manual muy documentado.

Amnesia

Herramienta desarrollada por OpenAire para la anonimización de datos personales a partir de datos tabulados.

Las usuarias y usuarios pueden filtrar los datos que se van a mostrar u ocultar determinados valores. Además, permite la creación de jerarquías de generalización para evitar la identificación de los datos a través de combinaciones únicas (ej.: ciudad de nacimiento y edad) a través de K-anonimidad.

La plataforma de Amnesia cuenta además con recursos y tutoriales.

Herramientas de análisis de datos basadas en código:

R Studio

R es un lenguaje de programación desarrollado en código abierto orientado a la computación estadística y a la visualización de análisis de datos. Cuenta con una aplicación específica, R Studio.

El lenguaje R es uno de los más usados en los campos de aprendizaje automático, minería de datos e inteligencia artificial, incluyendo áreas diversas como la investigación biomédica, el análisis estilométrico en lingüística computacional o las matemáticas financieras.

R dispone numerosas herramientas de análisis estadístico de datos, y cuenta con una comunidad de programadores que contribuyen al desarrollo de librerías y herramientas específicas. Una de las librerías más usadas en la creación de gráficos es ggplot2.

Por último, es importante destacar que R permite la creación de gráficos de alta calidad que pueden ser exportados en diversos formatos.

La principal desventaja del uso de R es su complicada curva de aprendizaje, pero existen numerosos recursos de aprendizaje como la guía para principiantes o esta guía de Fradejas Rueda.

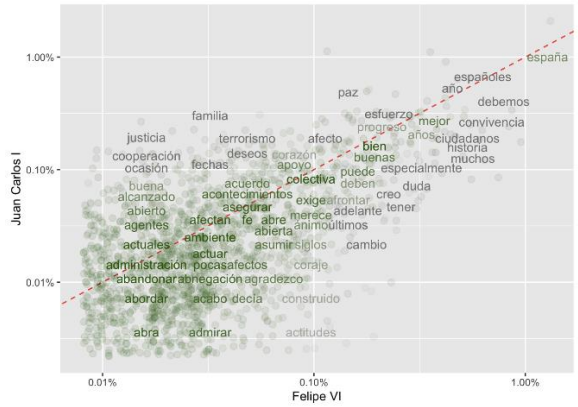


Gráfico creado con R-Studio. Fuente: Fradejas Rueda (2022)

Python

Python es un lenguaje de programación multiplataforma muy flexible. Se puede ejecutar en las terminales de código del ordenador o a través de la interfaz web Jupyter Notebooks.

Al igual que R, cuenta con numerosas librerías que incluyen herramientas para la creación de visualizaciones (Matplotlib), análisis matemático de datos (NumPy o Pandas) o el procesamiento de lenguaje natural (spaCy).

El lenguaje de Python es más sencillo que otros lenguajes de programación y su curva de aprendizaje es más moderada. Para comenzar a aprender, puede consultarse la guía LearnPython y para usar los Jupyter Notebooks la guía How to Use Jupyter Notebook.

Herramientas de visualización:

En esta sección se presenta una herramienta que está diseñada exclusivamente con este fin, pero algunas de las herramientas ya citadas como **R Studio** o **Python**, incluso otras herramientas más clásicas como las hojas de cálculo, permiten visualizar datos en gráficos y mapas.

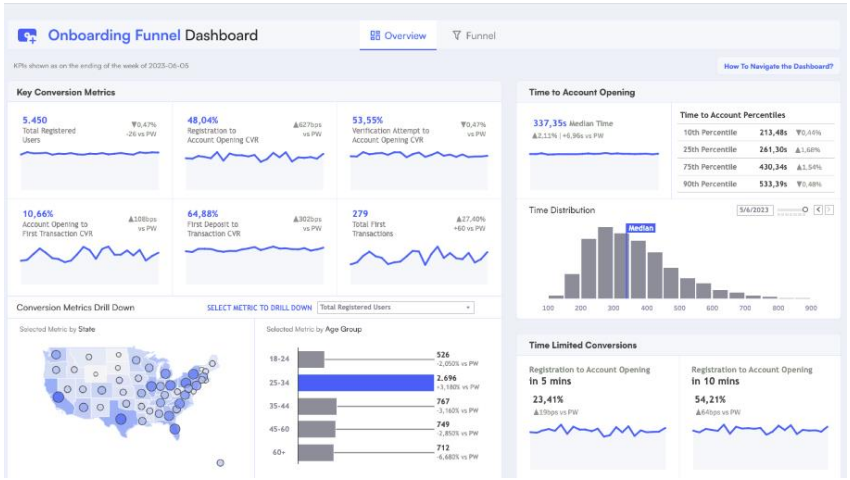
Tableau Public

Herramienta de Bussiness Intelligence y visualización de datos mediante gráficos y cuadros de mando. Es una herramienta de software propietario y cuenta con una versión gratuita (Tableau Public) y otra de pago con más funcionalidades (Tableau Desktop). La versión gratuita exige la publicación de los resultados en su página web, por lo que es necesario anonimizar cualquier dato sensible antes de trabajar en esta aplicación.

Tableau permite la creación de diferentes gráficos a partir de hojas de cálculo y bases de datos. Pueden realizarse alguna operación con los datos se pueden en la misma aplicación, así como crear nuevas variables, o transformar el tipo de datos.

Una de las principales ventajas de Tableau es la facilidad de uso, ya que no necesita de conocimientos en programación, y la publicación de gráficos y cuadros de mandos interactivos.

Para aprender a usar Tableau Public puedes consultarse sus tutoriales o explorar la galería para inspirarse en la creación de cuadros de mandos.



Ejemplo de cuadro de mando en Tableau Public. Fuente: #VOTD de Varun Jain He/Him

Sistemas de información geográfica (GIS):

Existen programas específicos para trabajar con datos que están geolocalizados, normalmente datos asociados a coordenadas geográficas o nombres administrativos. Estos programas**, llamados Sistemas de información de geográfica (GIS)** permiten procesar datos en grandes cantidades y generar mapas. Otras herramientas que trabajan con datos geográficos son **R Studio** o **Tableau Public.**

ArcGIS Online

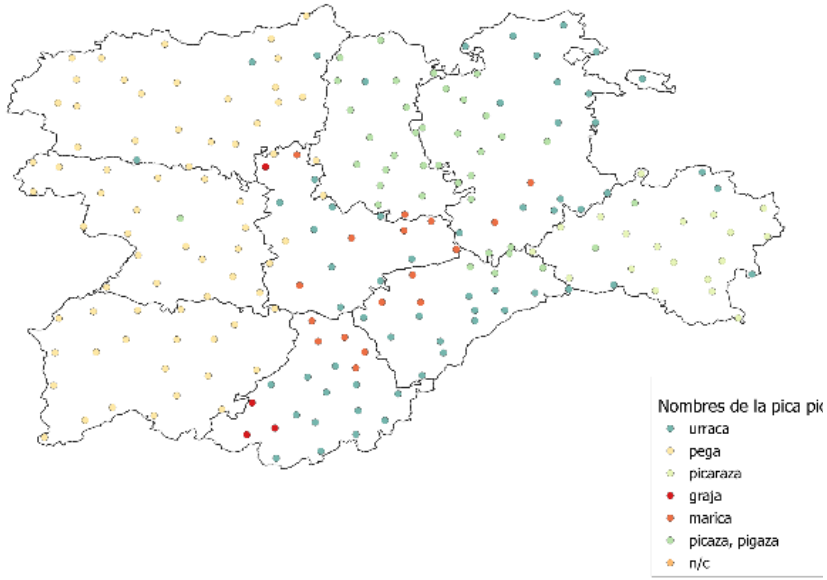
Herramienta para tratamiento y análisis de datos geográficos y creación de mapas.

Es una herramienta de software propietario y cuenta con una versión gratuita (ArcGIS Online) y otra de pago con más funcionalidades (ArcGIS Pro). ArcGIS Online está basada en la nube.

En su página pueden encontrarse numerosos recursos de formación.

QGIS

Sistema de información geográfica de software libre y código abierto. Ofrece funcionalidades similares a ArcGIS pero tiene una interfaz algo menos intuitiva, aunque cuenta con documentación de uso.



Mapa creado con QGIS en el marco del proyecto CORPAT

Depósito y publicación de datos

Repositorios multidisciplinares

Existen numerosos repositorios de datos para el depósito de *datasets.* Recuerda la importancia de completar adecuadamente los metadatos del *dataset* para que pueda ser localizado y reutilizado.

Entre los repositorios multidisciplinares, destacan:

e-cienciaDatos

Repositorio de datos del Consorcio Madroño, del que UNED forma parte.

Creado en 2016, e-cienciaDatos, es el repositorio de datos de carácter multidisciplinar que alberga los conjuntos de datos científicos finales gestionados por la comunidad investigadora de las universidades públicas de la Comunidad de Madrid y la UNED, que son miembros del Consorcio Madroño, con el fin de dar visibilidad a dichos datos, garantizar su preservación y facilitar su acceso y reutilización.

E-cienciaDatos utiliza software libre del proyecto Dataverse, desarrollado por la Universidad de Harvard que permite compartir, preservar, citar, explorar y analizar datos de investigación. También implementa el protocolo OAI-PMH para la recolección de datos e interoperabilidad.

Los datasets se describen conforme a uno o varios estándares de metadatos como Dublin Core y DataCite, para garantizar la compatibilidad con la infraestructura europea de investigación OpenAIRE. También se asignan identificadores de objetos digitales (DOis) a los datasets y a los ficheros.

Cada dataset debe ir acompañado de un fichero de texto en formato txt (texto plano) que, bajo el título de Readme, explique en detalle la metodología empleada para la recogida o generación de datos y el software para su uso, además de información general, autoría, palabras clave, cobertura geográfica, cobertura temporal y licencia de uso.

El depósito de los datos se hace mediante archivo delegado. Son las bibliotecas de cada universidad las que se ocupan de hacerlo, por lo que el personal investigador deberá solicitárselo a su biblioteca.

Se incluyen licencias libres para los datasets: CC0, CC-BY, CC-BY-SA y Open Data Commons Public Domain Dedication and License (PDDL)

E-cienciaDatos ha obtenido el prestigioso certificado CoreTrustSeal, que garantiza el cumplimiento de los requisitos exigidos en el programa Horizonte Europa.

Sus datos se recolectan y están accesibles en: OpenAire Explore, Harvard Dataverse Network y Google Dataset Search.

En resumen, e-cienciaDatos permite al personal investigador:

- Asociar el identificador ORCID al conjunto de datos.
- Enlazar la publicación y los datos que sustentan la investigación.
- Aumentar la visibilidad de la investigación mediante la asignación de dois a los datasets.
- Cumplir con las exigencias de los programas de financiación.
- Responder a las demandas de las revistas relativas a la disponibilidad de los datos que sustentan los resultados de la investigación publicada.

Zenodo

Repositorio multidisciplinar de datos de la comisión europea, a través del proyecto OpenAire. Permite el depósito de conjuntos de datos, así como otros muchos tipos de resultados de investigación.

Buscador de repositorios

Además de los repositorios multidisciplinares, hay áreas temáticas específicas que usan sus propios repositorios. Para encontrar repositorios temáticos, se recomienda el uso de Re3data.

Re3data

Registro de repositorios de datos de investigación que permite el descubrimiento de repositorios de datos temáticos.

Organización de los datos

Para que los *datasets* depositados en repositorios de datos cumplan los principios FAIR, es necesario que los datos estén organizados y acompañados de una serie de documentos que ayuden a interpretar y reutilizar estos datos.

Para organizar y documentar los datos correctamente es imprescindible seguir estas pautas:

- **Utiliza formatos adecuados.** Usa formatos preferiblemente abiertos y no propietarios que aseguren la vida a largo plazo de los datos sin pérdida de información. Puedes ver una lista de formatos recomendados en la biblioguía Gestión y depósito de datos científicos de la Biblioteca UNED.
- **Cuida la nomenclatura y estructura de los archivos.** Una buena estructura de los ficheros, con nombres coherentes y una jerarquía clara, permite a cualquier persona encontrar e interpretar los datos de manera rápida y precisa.
- **Almacena los datos de forma segura.** Realiza copias de seguridad y protege los datos sensibles.
- **Describe los datos de investigación**. Los metadatos, junto a la documentación que acompaña al *dataset*, permiten localizar e interpretar los datos de forma correcta.

Estructura del *dataset*

Para crear una organización de datos jerárquica y lógica, que sea fácil de interpretar por cualquiera, puedes seguir las siguientes recomendaciones:

- Organiza ficheros y carpetas de forma sistemática en una estructura de niveles.
- Evita estructuras con más de tres niveles de carpetas.
- Unifica la nomenclatura de los ficheros para que sea clara y concisa.
- Establece un sistema claro para registrar cambios y versiones.
- Elimina los ficheros que no interesen.

Para más información, puedes consultar la guía de la Cornell University consejos para organizar los datos.

Nomenclatura de ficheros

Al igual que la estructura, los nombres de los archivos y carpetas es crucial para asegurar la recuperación y reinterpretación de los datos.

El elemento principal del nombre del fichero ha de ser la descripción del contenido del mismo. Además, puede incluirse una serie de elementos si son necesarios para su identificación:

- Número de versión
- Fecha de creación
- Nombre del creador
- Nombre del equipo de investigación / departamento asociado con los datos
- Fecha de publicación
- Número del proyecto

Los nombres de los ficheros han de ser concisos y explicativos, es decir, ni muy largos ni muy cortos, presentando toda la información necesaria para la identificación e interpretación del fichero.

Además, es importante seguir una serie de pautas:

- Utiliza nombres cortos y relevantes.
- Evita los caracteres especiales: ~ ¡ ! @ # $ % ^ & * ( ) ` ; < > ¿ ? , [ ] { } ' " |
- Usa guiones bajos (snake_case) en lugar de espacios o cambios mayúsculas-minúsculas (camelCase).

- Crea pautas para fechas y versiones.

Un ejemplo de un archivo bien nombrado sería:

Encuestas_Sindicatos_20230723_V1. csv

Si necesitas renombrar ficheros en masa, existen herramientas como Bulk Rename Utility o Renamer.

Desde la planificación del proyecto, es necesario establecer una serie de pautas para nombrar archivos de manera sistemática, sobre todo en lo referente a **fechas** y **versiones.**

Para las **fechas**, se recomienda usar los estándares ISO 8601 que propone el sistema YYYYMMDD, de tal modo que los archivos queden ordenados cronológicamente. Con este sistema, la fecha 02/12/2022 quedaría 20221202.

En cuanto a las versiones, es imprescindible evitar etiquetas como "revisión", "final", "definitiva", etc. En cambio, se recomienda identificar las versiones con números ordinales y los pequeños cambios con números decimales. Por ejemplo: v1.1, v1.2, v2…

Puedes encontrar más recomendaciones sobre la nomenclatura de los datos y variables en la guía Gestión de Datos de Investigación de la Universidad Pablo de Olavide.

Metadatos

Los metadatos, los datos sobre los datos, son información estructurada que permiten identificar y definir los datos.

Existen numerosos tipos de metadatos, desde el nombre de los archivos, o los datos que acompañan al *dataset* en el repositorio. Además de estos, hay ocasiones en las que podemos necesitar usar un **esquema de metadatos** que se adapte a nuestra área de conocimiento, asegurando un mayor impacto y visibilidad de nuestra investigación.

Un esquema de metadatos es un conjunto de metadatos definido. Cada uno de los elementos cuenta con un nombre y definición específicos, se utilizan siguiendo reglas sintácticas y un vocabulario controlado.

Existen numerosos esquemas de metadatos y cada disciplina tiene su propio estándar. Digital Curation Center (DCC) dispone de un directorio por disciplinas para metadatos de investigación.

Documentación de los datos

Es necesario acompañar a los datasets con una serie de documentos que ayuden a interpretar y reutilizar los datos de investigación. Estos documentos varían según el tipo de datos o proyectos, pero suelen incluir ficheros readme, diccionarios de datos y licencias.

Estos ficheros pueden referirse al conjunto de todos los datos, o a una pequeña parte del dataset.

Fichero *readme*

El principal documento asociado a los *datasets* tiene la forma de fichero *readme.*

Los ficheros *readme* se realizan en archivos de texto plano TXT. Normalmente, están escritos en inglés e incluyen al menos la siguiente información:

- Contexto: información del proyecto, identificación de autores
- Información de acceso: licencia de los datos
- Estructura y archivos: lista de carpetas y archivos y sus relaciones
- Metodología de recopilación, tratamiento y análisis de datos
- Cambios y versiones

El Consorcio Madroño tiene una plantilla de fichero readme. Por su parte, la Cornell University ha creado una guía con consejos para redactar ficheros readme que incluye un modelo descargable.

Diccionario de datos

Además del fichero *readme*, o si procede dentro del mismo, es recomendable realizar un diccionario de datos, que define de manera univoca las variables del fichero para evitar interpretaciones erróneas.

Los diccionarios de datos pueden tener la forma archivo tabular, PDF o JSON, entre otros, e incluye la siguiente información:

- Nombres y definiciones de cada variable
- Tipo de datos (números enteros, cadenas de texto, fechas, etc.)
- Longitud del campo
- Campo requerido
- Posibilidad de tener valores nulos
- Explicación de cualquier código o sistema empleado en los datos

Puedes descargar este ejemplo de diccionario de datos para ver cómo se estructuran estos documentos.

Licencia

La licencia de los datos se incluye normalmente en el fichero *readme*, aunque también puede explicitarse en un documento aparte o en el registro del repositorio. OpenAire recomienda usar las licencias más abiertas posibles para asegurar que los principios de Ciencia Abierta y evitar cualquier obstáculo en el acceso al conocimiento.

La comunidad científica utiliza mayoritariamente las Creative Commons, que especifican los usos permitidos a través de una combinación sencilla de licencias. Las licencias más abiertas son CC-0, equivalente al dominio público, o en su defecto Creative Commons CC-BY, que requiere la atribución de autoría.

Licencia CC-0

Los datos pueden no tienen derechos reservados. Pueden usarse o modificarse con cualquier fin y no es necesario atribuir la autoría

Licencia CC-BY

La única limitación es la atribución de la autoría (BY). Los datos pueden usarse y adaptarse con cualquier fin.

Otras licencias usadas comúnmente son las licencias Open Data Commons, de la Open Knowledge Foundation, que están diseñadas específicamente para bases de datos y datos abiertos.

En este caso, la licencia equivalente a CC-0 sería la licencia Public Domain Dedication and License (PDDL), aunque también puede exigirse la atribución de la autoría con ODC-By.

Otros recursos

Consorcio Madroño (2022). *Cómo depositar sus datos de investigación en e-cienciaDatos.* http://www.consorciomadrono.es/docs/guia_gestion_datos_investigacion_2022-01-04.pdf

Cornell University. *Guide to writing "readme" style metadata* Recuperado en 2023 de: https://data.research.cornell.edu/content/readme

Cornell University. *Research Data Management: File Organization*. Recuperado en 2023 de: https://simmons.libguides.com/c.php?g=814790&p=5983200

Fradejas Rueda, J. (2022). *Cuentapalabras. Estilometría y análisis de texto con R para filólogos.* http://www.aic.uva.es/cuentapalabras/

Iniciativa Aporta (2021). *Herramientas de procesado y visualización de datos* Ministerio de asuntos económicos y transformación digital

Inicciativa Aporta (). Guía práctica para la publicación de datos tabulares en archivos CSV. https://datos.gob.es/es/documentacion/guia-practica-para-la-publicacion-de-datos-tabulares-en-archivos-csv

MANTRA. Research Data Training. Recuperado en 2023 de: https://mantra.ed.ac.uk/

R4DS (2017). *R for Data Science*. https://r4ds.had.co.nz/

The R-Graph Gallery. Recuperado en 2023 de: https://r-graph-gallery.com/all-graphs

UNED. Gestión y depósito de datos científicos: Inicio. Recuperado en 2023 de: https://uned.libguides.com/gestion_datos_cientificos

Universidad Pablo de Olavide. Gestión de datos de investigación. Recuperado en 2023 de: https://guiasbib.upo.es/gestion_datos_de_investigacion/

Universidad Pablo de Olavide. *Introducción a la visualización de datos de investigación.* Recuperado en 2023 de: https://guiasbib.upo.es/visualizacion_datos_investigacion_vpublica

University of Leeds. *Research data management explained.* Recuperado en 2023 de: https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained

University of Wisconsin-Madison (2023). Data Visualization in R with ggplot2. https://ssc.wisc.edu/sscc/pubs/dvr/index.html

University of Wisconsin-Madison (2023). Data Wrangling with R. https://sscc.wisc.edu/sscc/pubs/dwr/index.html